Nucleic Acids Research, 2007, Vol. 35, Database issue ***D511–D514***
doi:10.1093/nar/gkl972

BRENDA, AMENDA and FRENDA: the enzyme information system in 2007

Jens Barthelmes, Christian Ebeling, Antje Chang, Ida Schomburg and Dietmar Schomburg*

Institute of Biochemistry, University of Cologne, Zülpicher Strasse 47, 50674 Köln, Germany

Received September 15, 2006; Revised and Accepted October 20, 2006

ABSTRACT

The BRENDA (BRaunschweig ENzyme DAtabase) enzyme information system (http://www.brenda.uni-koeln.de) is the largest publicly available enzyme information system worldwide. The major parts of its contents are manually extracted from primary literature. It is not restricted to specific groups of enzymes, but includes information on all identified enzymes irrespective of the enzyme's source. The range of data encompasses functional, structural, sequence, localisation, disease-related, isolation, stability information on enzyme and ligand-related data. Each single entry is linked to the enzyme source and to a literature reference. Recently the data repository was complemented by text-mining data in AMENDA (Automatic Mining of ENzyme DAta) and FRENDA (Full Reference ENzyme DAta). A genome browser, membrane protein prediction and full-text search capacities were added. The newly implemented web service provides instant access to the data for programmers via a SOAP (Simple Object Access Protocol) interface. The BRENDA data can be downloaded in the form of a text file from the beginning of 2007.

INTRODUCTION

The BRENDA (BRaunschweig ENzyme DAtabase) enzyme information system (1,2) is a manually annotated repository for enzyme data. Originally intended and published as a series of books (3) in 1987, it was transformed into a publicly available database in 1998 and has been curated and continuously improved at the University of Cologne since then. Its contents are not restricted to specific groups of enzymes, but include information on all enzymes that have been classified in the EC scheme of the IUBMB (International Union of Biochemistry and Molecular Biology) irrespective of the enzyme's source. The range of data includes the catalyzed reaction, detailed description of the substrate, cofactor and inhibitor specificity, kinetic data, structure properties, information on purification and crystallization, properties of mutant enzymes, participation in diseases and amino acid sequences. Each single entry is linked to the enzyme source (organism and, if applicable, the tissue and/or the protein sequence) and to the literature reference. Data queries can be performed by a number of different ways, including an EC-tree browser, a taxonomy-tree browser, an ontology browser and a combination query of up to 20 parameters. The newly implemented web-service provides instant access to the data for programmers via a SOAP interface (Simple Object Access Protocol).

CONTENTS OF BRENDA

Data statistics

In summary BRENDA contains ~1.3 million manually annotated data, on average 300 single entries per EC-number (Table 1). Enzymes of 7500 different organisms are covered. With ~170 000 single data human enzymes are the most thoroughly described in the literature (Figure 1), followed by enzymes of *Rattus norvegicus* (~132 000 entries) and *Escherichia coli* (~93 000 entries)

New information fields

pI value: The isoelectric point is now included. This value is of significance for the purification procedure allowing conclusions about the solubility of the enzyme and its motility in electrophoretic procedures.

K_i value: 14 014 inhibition constants are presently included in the database. Each value is connected to the enzyme, to the inhibitor and, where available, to the 2D structure of the molecule.

Engineered enzymes: The reactivity of mutant enzymes can reveal detailed insight into the catalytic process and may give valuable clues about the active sites, the mechanism of the reaction or the regulation. Meanwhile ~19 000 engineered enzymes are described in the database. For each single modification of the protein sequence, the properties of the resulting enzyme are described. Kinetic data for these enzymes are included in the respective database sections.

MOLECULAR STRUCTURE-BASED QUERIES

When searching for molecules which interact with the enzyme (substrates, products, cofactors, inhibitors, activating substances, etc.) different query procedures are possible.

*To whom correspondence should be addressed. Tel: +49 221 470 6440; Fax: +49 221 470 5092; Email: D.Schomburg@uni-koeln.de

- *Using the name of the compound*: This option returns not only the data stored for the ligand under the given name but applies the integral molecular thesaurus based on the INChI (IUPAC International Chemical Identifier) (4) codes of ~53 000 molecular structures stored as molfiles.
- *Performing a substructure search with the integrated JME Editor (5)*: The result page of this function displays the images, names and synonyms of the found compound, their function when interacting with the enzyme and also provides a button for an immediate BRENDA search.

ONTOLOGIES

The BRENDA Ontologies section allows to search in all publicly available ontologies of biochemical, anatomic, developmental, chemical and medical terms such as Gene Ontology (6) or MeSH (7) published in open biomedical ontology format (http://obo.sourceforge.net). If possible, terms are cross-linked to other ontologies and BRENDA enzyme data. The use of umbrella terms allows to search, for example, for complete classes of chemical compounds in the BRENDA database.

Table 1. Data statistics for the various sections of the database

Enzyme information	Single data[a]
Names and synonyms	70 972
Isolation and preparation	53 364
Stability	34 532
Reaction and specificity	396 760
Enzyme structure	232 824
Functional and kinetic parameters	191 134
Organism-related information	80 964
References	91 403
Enzyme application	3854
Enzyme-related diseases	52 558
Mutant enzymes	18 194

[a]These numbers refer to the combination of enzyme-organism-value.

NEW DATABASES AT THE BRENDA HOST

FRENDA

FRENDA (Full Reference ENzyme DAta) is an additional database to BRENDA available to the academic community with BRENDA release 6.2 (June 2006). FRENDA aims at providing an exhaustive collection of indexed literature references containing organism-specific enzyme information. Compared to a standard PubMed (8) query, FRENDA returns also all references on the enzyme published under one of its synonyms.

FRENDA currently covers 1.4 million enzyme/organism combinations from 550 000 distinct references, automatically extracted from more than 16 million PubMed abstracts (June 2006) (8). The scientific articles are pre-filtered using MeSH terms—only references declared as 'enzyme' hits are used (1.6 million remaining abstracts). FRENDA uses a dictionary-based approach for recognizing named entities (enzymes, organisms) in titles and abstracts. The dictionaries are compiled from BRENDA and NCBI Taxonomy (8). In a two-step approach, references with enzyme hits in title, abstract or MeSH terms are searched for co-occurring organism names (scientific names and synonyms).

The results of this indexing process were classified into four reliability categories depending on the occurrence of

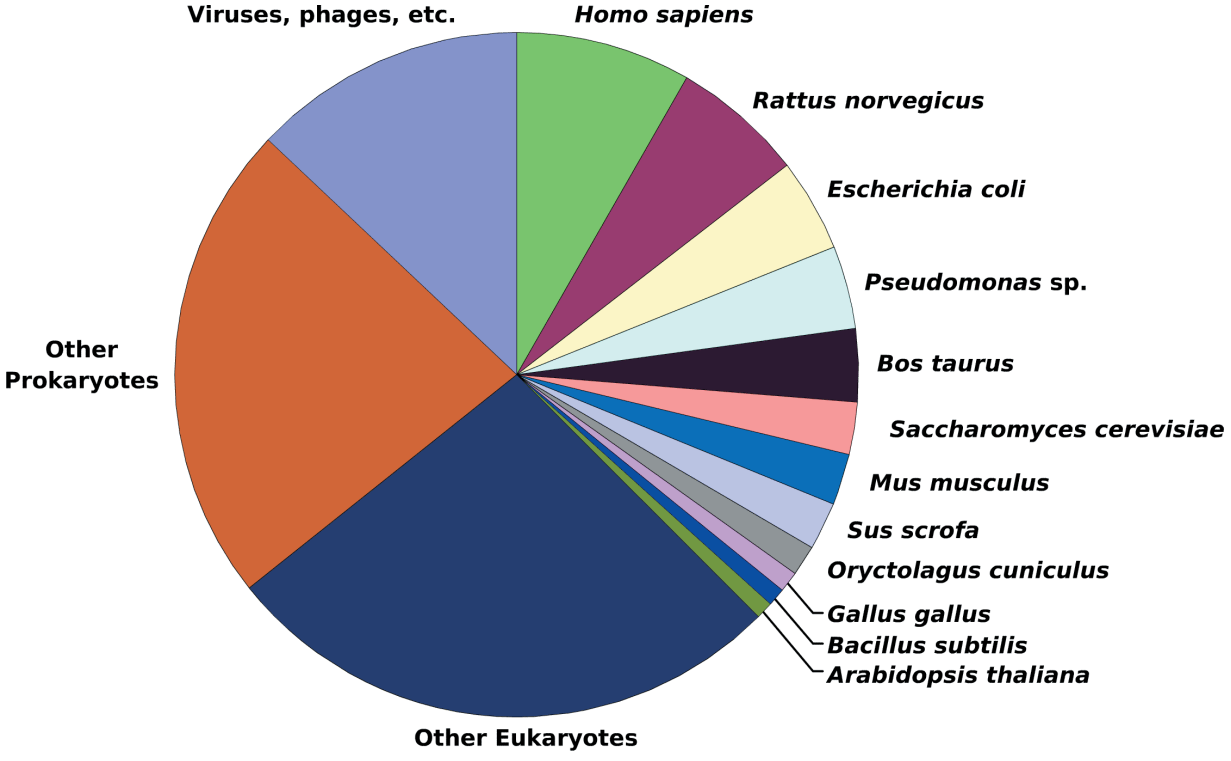

Figure 1. Organism coverage in BRENDA data.

search terms in title and/or abstract and/or MeSH terms. This classification is provided with the commentaries in the FRENDA database.

The manual evaluation of the quality of the FRENDA approach using 250 randomly chosen results indicates a precision of 64.8% with a recall of 72% from a set of 250 manually annotated enzyme-related literature references.

AMENDA

As a subset of FRENDA, AMENDA (Automatic Mining of ENzyme DAta) currently covers organism-specific information on enzyme localization (>30 000 records, compared with 17 000 records in BRENDA) and source tissues (~150 000 records, compared with 38 000 records in BRENDA) from a text-mining procedure (J. Barthelmes, C. Ebeling and D. Schomburg, unpublished data).

Search terms for enzyme names, organism names, localization and sources and tissues are compiled from BRENDA enzyme synonyms, the BRENDA tissue-tree (http://obo.sourceforge.net/cgi-bin/detail.cgi?brenda) and the NCBI Taxonomy (8). AMENDA is based on the FRENDA co-occurrence approach. Protozoa, viruses and bacteria are excluded for tissue search. References with enzyme/organism hits are searched for occurrences of tissue terms (singular and plural) and localization terms in title, abstract, and MeSH terms and further evaluated based on text-mining criteria.

The text-mining approach described above was tested on 200 randomly selected results. A precision of ~76.0% and a recall of 11.7% for the combined search terms enzyme-organism-tissue/localization was achieved. In a way similar to FRENDA, the commentaries indicate the individual reliability level for each data set.

BRENDA GENOME EXPLORER

The BRENDA Genome Explorer is an enzyme-centered genome visualization tool for browsing and comparing enzyme annotations in full genomes. It closes the gap between genomic and enzymatic data and allows the alignment of genomes at a given enzyme-coding gene and its orthologs, thus allowing to visually compare the genomic environment of the gene in different organisms (Figure 2). The underlying genome database is compiled from EBI Genomes (9) and ENSEMBL (10) and supplemented by UniProt (11) annotations. It can be searched for specific proteins via names, EC-numbers, or UniProt accessions, allowing for a highly target-oriented search.

TRANSMEMBRANE PROTEIN PREDICTION

Transmembrane helices for enzymes are predited with TMHMM (TransMembrane Hidden Markov Model) developed by Sonnhammer *et al.* (12). With the aid of this tool it is possible to predict the number, the size and the location of transmembrane helices, thereby discriminating soluble and membrane-bound enzymes.

ACCESSIBILITY

BRENDA is accessible via the various search options (quick search, advanced search, ontologies, sequence search, Genome Explorer, etc.). The database can be downloaded

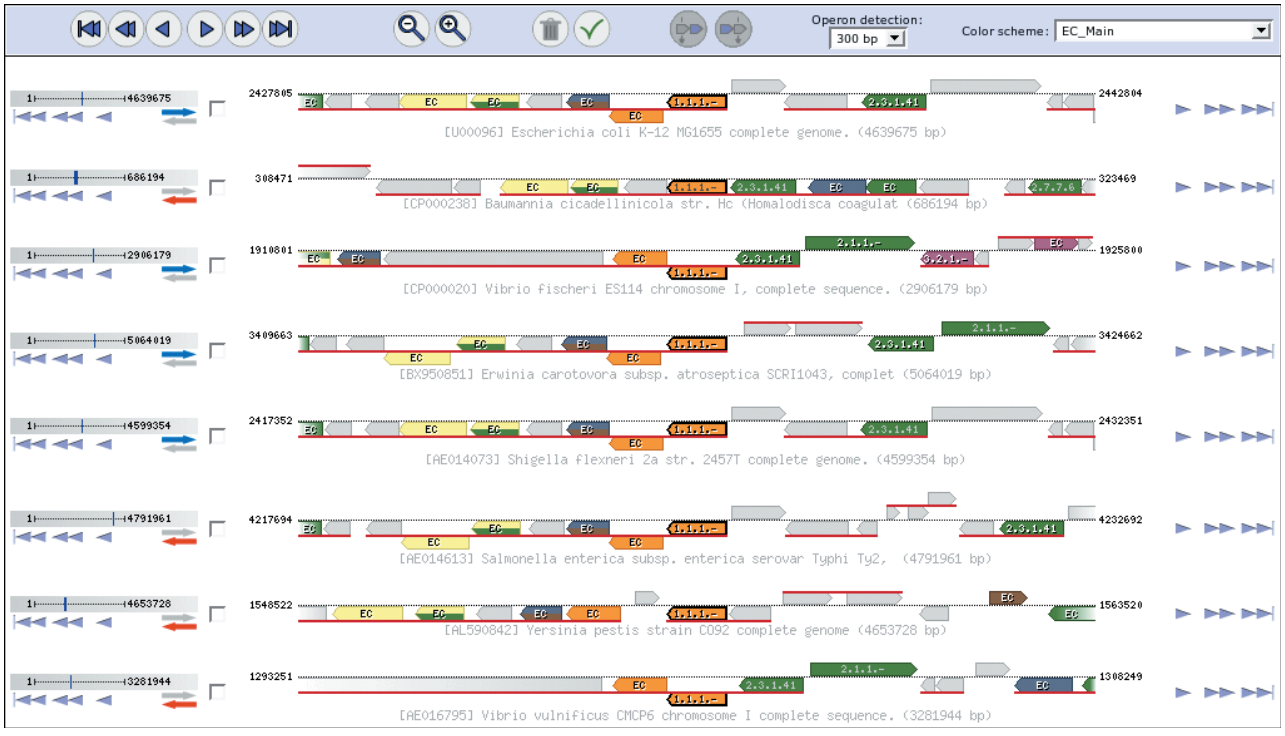

Figure 2. BRENDA Genome Explorer showing a genome alignment for *Escherichia coli* erythronate-4-phosphate dehydrogenase, EC 1.1.1.290 (excerpt).

as a text file. Access to AMENDA and FRENDA requires a registration.

SOAP-BASED WEB SERVICE

Web services provide a simple way to access the data collection without the need for downloading, parsing and preparing an entire database for local queries. Web services are independent of the internal organization of the database and avoid parsing problems caused by changes in the text file structure.

BRENDA now provides a SOAP (http://www.w3.org/TR/soap) based web service comprised of 148 methods covering 52 data fields. Flexible queries can be performed directly from programs written in different programming languages (Perl, Java, C++, Python, PHP) on data fields such as substrate, K_m-value and pH-optimum. For any given record returned, a set of complete literature references can be retrieved using unique reference identifiers. Every data field may be queried by providing at least one of the three parameters EC-number, organism, or—if applicable—ligand structure identifier. The ligand structure identifier, which can be queried with the name of a chemical compound, is used to ensure that all synonyms for a given molecular structure are also retrieved.

The BRENDA web service also gives access to the data using identifiers from other databases like UniProt (11) or NCBI Taxonomy (8), as well as ontologies like Gene Ontology (6) or BRENDA Tissue Ontology. The ontology-based search allows for queries based on entire branches of the hierarchy, avoiding a complex search for all leaves in the given branch. For example, an ontology-based search for the term 'brain' or the respective Gene Ontology identifier will return all tissues and cell types under the umbrella term 'brain'. The same method can also be applied to search for whole groups of organisms.

The documentation of the BRENDA web service including examples in different programming languages is available at http://www.brenda.uni-koeln.de/soap.

CONCLUSIONS

In the past year the BRENDA enzyme information system has made a big step forward not only by a formidable increase in the annotation speed but also by inclusion of data based on text-mining approaches and by the development of different new methods for data access. The new funding by an EU grant allows to increase the annotation speed even further to bring the backlog down to less than one year and will also allow to substantially increase the percentage of ligands with full structural information.

ACKNOWLEDGEMENTS

This work was supported by the European Union (FELICS, Free European Life-Science Information and Computational Services): 0121902 (RII3) and the German BMBF, Federal Ministry of Education, Science, Research and Technology (NGFN, Nationales Genomforschungsnetz: 0313398B) Funding to pay the Open Access publication charges for this article was provided by the German 'Bundesministerium für Bildung und Forschung (BMBF)'.

Conflict of interest statement. None declared.

REFERENCES

1. Schomburg,I., Chang,A., Ebeling,C., Gremse,M., Heldt,C., Huhn,G. and Schomburg,D. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res*., **32**, D431–D433.
2. Schomburg,I., Chang,A. and Schomburg,D. (2002) BRENDA, enzyme data and metabolic information. *Nucleic Acids Res*., **30**, 47–49.
3. Schomburg,D. and Schomburg,I. (2001) *Springer Handbook of Enzymes*, 2nd edn. Springer, Heidelberg, Germany.
4. Stein,S.E., Heller,S.R. and Tchekhovski,D. (2003) An open standard for chemical structure representation—the IUPAC chemical identifier. *Nimes International Chemical Information Conference Proceedings*, Nimes, France, pp. 131–143.
5. Ertl,P. and Jacob,O. (1997) WWW-based chemical information system. *Theochemistry*, **419**, 113–119.
6. Gene Ontology Consortium. (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res*., **34**, D322–D326.
7. National Library of Medicine. (1960) *Medical Subject Headings: Main Headings, Subheadings, and Cross References Used in the Index Medicus and the National Library of Medicine Catalog, 1st edn.* Washington, DC: U.S. Department of Health, Education, and Welfare.
8. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al*. (2006) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **34**, D173–D180.
9. Cochrane,G., Aldebert,P., Althorpe,N., Andersson,M., Baker,W., Baldwin,A., Bates,K., Bhattacharyya,S., Browne,P., van den Broek,A. *et al*. (2006) EMBL Nucleotide Sequence Database: developments in 2005. *Nucleic Acids Res*., **34**, D10–D15.
10. Birney,E., Andrews,D., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cox,T., Cunningham,F., Curwen,V., Cutts,T. *et al*. (2006) Ensembl 2006. *Nucleic Acids Res*., **34**, D556–D561.
11. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al*. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*., **34**, D187–D191.
12. Sonnhammer,E.l.L., von Heijne,G. and Krogh,A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. In Glasgow,J., Littlejohn,T., Major,F., Lathrop,R., Sankoff,D. and Sensen,C. (eds), *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, CA, pp. 175–182.